SPR Day 3

Readings

PRML 1.2.3 Bayesian probabilities

2.1 Binary rand. vars
+ Beta distrib.
+ Gamma functions

Goals

Review continuous rand. variables

Introduce Beta distribution
- PDF formula
- Graphs of PDF
- What is a Gamma function?

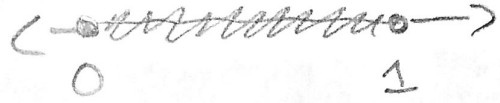
Beta-Bernoulli joint model for coin toss
- posterior distribution
- MAP estimation
- predictive distribution & estimation
- evidence

Continuous Random Variables

Examples

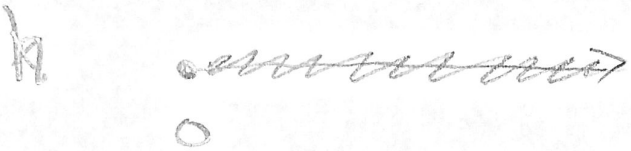
Probability coin ends up heads

μ sample space $\mu \in [0,1]$



Height of avg. student

h $h \in [0, \infty)$



Distance of random location from Medford

d $d \in (-\infty, +\infty)$

Sample spaces are continuous intervals of real line

Can talk coherently about the probability of interval events

what's probability that height is above 75 cm

$$p(H \geq 75)$$

Need different language for specific events

should not say what is probability height equals 75

instead, think about limits

$$pdf(H=75) = p(H \in [75, 75+\delta]) \text{ as } \delta \to 0$$

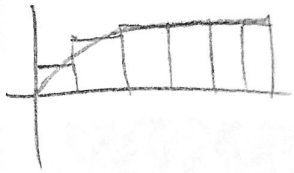
probability density function of random variable H

Two rules for a PDF function for continuous r.v. X

$$pdf(x) \geq 0 \quad \text{for all outcomes } x \in \mathcal{X}$$

$$\int_{x \in \mathcal{X}} pdf(x)\,dx = 1$$

Think of the Riemannian sum of this integral



$$\lim_{\Delta x \to 0} \sum_{n=1}^{N} pdf(x_n)\,\Delta x_n = 1$$

density of each bin | volume of each bin | mass

$$\text{density} = \frac{\text{mass}}{\text{volume}}$$

Can density be larger than one? Yes, if mass of bin is small enough

Joint probabilities also obey these rules

$X \in \mathbb{R}$, $Y \in \mathbb{R}$

$$pdf(x, y) \geq 0 \quad \text{and} \quad \iint pdf(x,y)\,dx\,dy = 1$$

Sum Rule

$$pdf(x) = \int_{y \in Y} pdf(x,y)\,dy$$

Product Rule

$$pdf(x,y) = pdf(x|y)\,pdf(y)$$
$$= pdf(y|x)\,pdf(x)$$

Beta distribution

Random variable μ

Sample space $\mu \in [0, 1]$

PDF function parameters $a > 0$, $b > 0$

$$\text{BetaPDF}(\mu | a, b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{const. wrt } \mu} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{\text{"interesting part" is a function of } \mu}$$

const. wrt μ: we'll call this "normalizing" constant

"interesting part" is a function of μ: $f(\mu | a, b)$

Gamma function: see notebook online

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

$\Gamma : \mathbb{R} \to \mathbb{R}_+$

scalar real $\Rightarrow$ scalar real

Remember:

$$\int_0^1 \text{BetaPDF}(\mu | a, b) \, d\mu = 1$$

$$\int_0^1 c(a, b) \, f(\mu | a, b) \, d\mu = 1$$

implies

$$\int_0^1 f(\mu | a, b) \, d\mu = \frac{1}{c(a, b)}$$

thus,

$$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} \, d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The Beta-Bernoulli Joint Model

We have defined a complete model (a joint distribution)

$$p(x_1, x_2, \cdots x_N, \mu) = \left[\prod_{n=1}^{N} \text{BernPMF}(x_n|\mu)\right] \cdot \text{BetaPDF}(\mu|a,b)$$

Bernoulli "likelihood" — Beta "prior"

Given this joint distribution, we can ask about related probability distributions that can be derived from this joint

"Posterior"

$$p(\mu|x_1 \cdots x_N) = \frac{p(x_1 \cdots x_N|\mu)\, p(\mu)}{p(x_1 \cdots x_N)}$$

("likelihood", "prior", "evidence")

via Bayes rule

"Evidence"

$$p(x_1 \cdots x_N) = \int_{\mu=0}^{1} p(x_1 \cdots x_N, \mu)\, d\mu$$

via sum rule

"Predictive Posterior"

$$p(x_N|x_1 \cdots x_{N-1}) = \int_{\mu=0}^{1} p(x_N|\mu)\, p(\mu|x_1 \cdots x_{N-1})\, d\mu$$

(likelihood, posterior)

via sum rule & Bayes rule

Posterior of μ for Beta-Bern model

Using Bayes rule:

$$p(\mu | x_1 \cdots x_N) = \left[\prod_{n=1}^{N} \text{BernPMF}(x_n|\mu)\right] \cdot \text{BetaPDF}(\mu|a,b)$$

$$\frac{}{p(x_1, \cdots x_N)}$$

$$= \frac{1}{p(x_1 \cdots x_N)} \cdot \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \cdot c(a,b) \cdot \mu^{a-1}(1-\mu)^{b-1}$$

Gather constant terms wrt μ

$$= \text{Const} \cdot \mu^{(\sum_n x_n)+a-1} \cdot (1-\mu)^{\sum_n (1-x_n)+b-1}$$

Why? $\left(\mu^{x_1} \cdot \mu^{x_2} \cdot \mu^{x_3} \cdots = \mu^{x_1+x_2+x_3+\cdots} = \mu^{\sum_n x_n}\right)$

$$= \text{const} \cdot \mu^{\boxed{s(x)+a}-1}(1-\mu)^{\boxed{r(x)+b}-1}$$

$$= c(a',b') \cdot f(\mu, a', b') \quad \text{where } s(x) = \sum_n x_n$$

$$r(x) = N - s(x)$$

We have written the posterior PDF so it looks like

$$\underset{\text{wrt } \mu}{\text{constant}} \cdot \mu^{a'-1} \cdot (1-\mu)^{b'-1} \quad \text{"} = \text{Beta} = \text{"}$$

Since PDF must equal one when integrated, we have:

$$\frac{1}{c(a',b')} = \int_0^1 \mu^{a'-1}(1-\mu)^{b'-1} d\mu$$

$$c(a',b') = \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}$$

Punchline:
Posterior is Beta with parameters

$a' = s(x) + a = \sum_n x_n + a$

$b' = r(x) + b = \sum_n 1 - x_n + b$

Towards MAP Estimator for Beta-Bernoulli

$$P(x_1, x_2 \cdots x_N, \mu) = \text{BetaPDF}(\mu|a,b) \prod_{n=1}^{N} \text{Bern}(x_n|\mu)$$

$x_1, x_2 \cdots x_N$: discrete; μ: continuous

We observe $x_1 \cdots x_N$, how to estimate μ? how to predict next word x_{N+1}?

several ways:

(1) find μ that maximizes $P(\mu | x_1 \cdots x_N)$
called maximum a-posteriori (MAP)

(2) take a more Bayesian approach (next time)

(1) MAP estimator

$$\mu^* = \arg\max_{\mu \in [0,1]} P(\mu | x_1 \cdots x_N)$$

$$= \arg\max_{\mu \in [0,1]} \frac{P(\mu) P(x_1 \cdots x_N | \mu)}{P(x_1 \cdots x_N)}$$

$P(x_1 \cdots x_N)$: const wrt μ

$P(x_1 \cdots x_N) = \int_\mu P(\mu, x) d\mu$

$$= \arg\max_{\mu \in [0,1]} \log \text{BetaPDF}(\mu|a,b) + \sum_n \log \text{BernPMF}(x_n|\mu)$$

$$= \arg\max_{\mu \in [0,1]} (a-1)\log\mu + (b-1)\log(1-\mu) + s(x)\log\mu + r(x)\log 1-\mu$$

MAP estimator (cont'd)

$$\mu^* = \underset{\mu \in [0,1]}{\arg\max} \overbrace{(s(x)+a-1)}^{s'} \log \mu + \overbrace{(r(x)+b-1)}^{r'} \log(1-\mu)$$

We've solved this before! See derivation of ML estimator

$$\mu^* = \frac{s'}{s'+r'} \quad \text{when} \quad s' \geq 0, \; r' \geq 0$$

$$\mu^* = \frac{s(x)+a-1}{\underbrace{N+a+b-2}_{= s(x)+r(x)}}$$

requires $a \geq 1$, $b \geq 1$

otherwise MAP does not exist

Interpretation

a acts like a "pseudocount" for heads

b acts like a "pseudocount" for tails

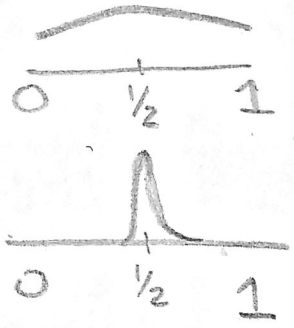
Problems w/ MAP estimator

- only works for $a>1, b>1$

  What if I have have $N=0$, $a=\frac{1}{2}, b=\frac{1}{2}$
  can't make predictions.

- Should I really condense a whole distribution down to just one estimate?

(a) (b)



should we act the same in these two cases?

Alternative

For predicting next coin (say x_{N+1} or x_*), consider:

ML $\quad p(x_{N+1} \mid \mu^{ML}) = \frac{s(x)}{N}$

MAP $\quad p(x_{N+1} \mid \mu^{MAP}) = \frac{s(x)+a-1}{N+a+b-2}$

condition on reductive estimates of data

Full Bayes $\quad p(x_{N+1} \mid x_1 \cdots x_N) = ? = \int p(x_{N+1} \mid \mu) p(\mu \mid x_1 \cdots x_N) d\mu$

condition on all data we observed

average over full distribution on μ

Posterior Predictive for next coin flip

Let's ask about proba. of heads:

$$p(x_{N+1}=1 \mid x_1 \cdots x_N) = \int \text{Bern}(x_{N+1}=1 \mid \mu)\, \text{Beta}(\mu \mid a', b')\, d\mu$$

$$= \int_0^1 \mu \, \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')} \mu^{a'-1} (1-\mu)^{b'-1} \, d\mu$$

$$= \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')} \int_0^1 \mu^{a'+1-1} (1-\mu)^{b'-1} \, d\mu$$

$\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')}$: const wrt μ, can be pulled out of integral

$\mu^{a'+1-1}(1-\mu)^{b'-1}$: combine like terms $= \mu^1 \cdot \mu^{a'-1}$

recognise as unnormalized Beta PDF

use identity about $c(a', b') = \int f(\mu, a', b')\, d\mu$

$$= \frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')} \cdot \frac{\Gamma(a'+1)\Gamma(b')}{\Gamma(a'+1+b')}$$

cancel out b' term

$$= \frac{\Gamma(a'+b')}{\Gamma(a')} \cdot \frac{a'\,\Gamma(a')}{(a'+b')\,\Gamma(a'+b')}$$

numerator: use identity $\Gamma(x+1) = x\,\Gamma(x)$ w/ $x = a$

denom.: use same w/ $x = a+b$

$$= \frac{a'}{a'+b'} = \boxed{\frac{s(x)+a}{N+a+b}}$$

This is the posterior predictive

works for all $a>0, b>0$